Capitolo VII

ATTENDIBILITA' DELLE STATISTICHE CAMPIONARIE

1. Distribuzione della media campionaria

Si può dimostrare matematicamente che *la distribuzione della media campionaria di un qualunque tipo di distribuzione teorica si approssima alla distribuzione normale man mano che aumenta la dimensione del campione*. In pratica questa proprietà è estremamente importante, perché essa permette l'uso delle proprietà della curva normale in tutti i casi in cui si considera il comportamento delle medie campionarie, *anche quando la variabile di base abbia una distribuzione non normale*. Va osservato, tuttavia, che *quando il campione è tratto da una popolazione che segue una distribuzione molto asimmetrica, per ottenere risultati accettabili è necessario che il campione sia di dimensione sufficientemente grande*.

Ogni campione di n elementi ha una propria media ed una propria varianza. A parità di dimensione del campione, le diverse combinazioni di n elementi che possono trarsi dalla popolazione di dati, sono dotate ognuna di una propria media e di una propria varianza. L'insieme delle medie dei campioni formati ognuno di n elementi sarà descritto da una distribuzione di frequenza che si chiama "*distribuzione della media campionaria nei campioni di n elementi*". Questa distribuzione, a sua volta, avrà una propria media ed una propria varianza. Inoltre vi saranno tante di queste distribuzioni quanti sono i valori che potrà assumere n, cioè la dimensione del campione.

Vediamo ora di determinare quali sono le principali caratteristiche della distribuzione delle medie campionarie, distinguendo tra i

campioni estratti da una popolazione finita e quelli che provengono da una popolazione infinita.

1.1 *Campioni estratti da una popolazione finita*

Ricerchiamo anzitutto la media e la deviazione standard della distribuzione formata dalle medie di tutti i campioni casuali possibili tratti da una popolazione finita, ricordando che la deviazione standard della distribuzione delle medie è chiamata "*errore standard della media*".

Si consideri una popolazione in cui ogni elemento sia indicato da x_i. Inoltre, N rappresenti il numero totale di elementi che costituiscono questa popolazione finita ed n indichi il numero di elementi contenuti in qualunque campione casuale tratto dalla stessa popolazione. Similmente M e σ rappresentino la media e la deviazione standard della popolazione, mentre m ed s rappresentino la media e la deviazione o errore standard del campione.

Il numero di tutti i possibili campioni casuali di n elementi (tutti diversi l'uno dall'altro) estraibili da una popolazione di N elementi in totale sarà numericamente pari a $C(N, n)$, cioè tanti quante sono le combinazioni di N elementi ad n ad n.

Ora si indichino con le lettere $m_1, m_2, m_3, \ldots\ldots\ldots, m_{C(N,n)}$, le medie di questi $C(N, n)$ campioni e si utilizzi il simbolo M_m per indicare il valore medio delle medie di tutti questi campioni.

Nelle condizioni or ora descritte, quindi, si avrà per esempio:

N° d'ordine dell'elemento
1° 2° 3°..... ..n°

$$m_1 = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

$$m_2 = \frac{x_3 + x_5 + x_9 + \ldots + x_k}{n}$$

.....

$$m_{C(N,n)} = \frac{x_i + x_s + x_t + \ldots + x_r}{n}$$

$$\sum_{i=1}^{i=C(N,n)} m_i = \frac{1}{n}\frac{n.C(N,n)}{N}(x_1 + x_2 + x_3 + \ldots + x_N)$$

Ogni m_i è calcolata su n termini ed in tutto si potranno calcolare $C(N, n)$ medie. Quindi nella somma di tutte le medie vi saranno $n.C(N, n)$ termini di tipo x_i ed alcuni di questi saranno uguali tra loro. Per calcolare quante volte una stessa x_i si presenta in tale somma, basta riflettere sul fatto che, d'altra parte, il numero delle x_i differenti l'una dall'altra è N. Di conseguenza una x particolare, diciamo ad esempio x_k, comparirà nella somma indicata $n.C(N, n)/N$ volte, per cui la somma di tutti i termini può essere scritta:

$$\sum_{i=1}^{i=C(N,n)} m_i = \frac{1}{n}\frac{n.C(N,n)}{N}(x_1 + x_2 + x_3 + \ldots + x_N)$$

Quindi sarà:

$$M_m = \frac{1}{C(N,n)}\sum_{i=1}^{i=C(N,n)} m_i = \frac{(x_1 + x_2 + x_3 + \ldots + x_N)}{N}$$

oppure:

$$M_m = \frac{1}{N}\sum_{i=1}^{i=N} x_i = M_x \qquad (1)$$

Dalla importante relazione che precede si trae la conclusione che *la media di tutte le medie campionarie è uguale alla media della popolazione*.

In modo simile si può calcolare l'errore standard (deviazione standard) della distribuzione delle medie campionarie. In questo caso i calcoli sono un poco più lunghi malgrado il ragionamento sia esattamente identico a quello fatto per il calcolo di M_m .

Se si indica con σ_m l'errore standard della media, si ha il risultato seguente:

$$\sigma_m = \sigma_x\sqrt{\frac{N-n}{n.(N-1)}} \qquad (2)$$

oppure, prendendo il quadrato di ambedue i lati della relazione, si ha:

$$\sigma_m^2 = \sigma_x^2\frac{N-n}{n.(N-1)}$$

$$\mu_m = \mu_x\frac{N-n}{n.(N-1)}$$

Come già detto, *queste formule sono valide nel caso di popolazioni finite*.

1.2 *Campioni estratti da una popolazione infinita*

La media della distribuzione campionaria delle medie relative a campioni che provengono da una popolazione infinita coincide con la media della popolazione da cui i campioni sono estratti anche quando la popolazione di origine sia formata da un numero infinito di unità.

Invece, quando la popolazione da cui sono estratti i campioni casuali è infinitamente grande, l'errore standard della media differisce da quello che si ottiene quando la popolazione di origine è finita. Infatti, quando la popolazione di provenienza è formata da un numero infinito di elementi, il valore di σ_m sarà uguale al limite dell'espressione (2), al tendere di N all'infinito. Quindi, si avrà che:

$$\lim_{N\to+\infty}\sigma_m = \lim_{N\to+\infty}\sigma_x\sqrt{\frac{N-n}{n.(N-1)}} = \sigma_x\lim_{N\to+\infty}\sqrt{\frac{1-\frac{n}{N}}{n.\left(1-\frac{1}{N}\right)}} = \sigma_x\sqrt{\frac{1}{n}}$$

ossia sarà:

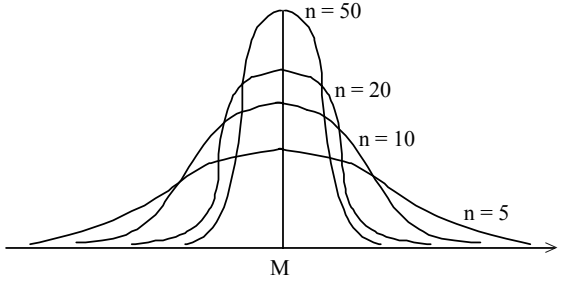
$$\sigma_m = \sigma_x\sqrt{\frac{1}{n}} = \sigma_x\frac{1}{\sqrt{n}} = \frac{\sigma_x}{\sqrt{n}}$$

$$\sigma_m^2 = \mu_m = \frac{\sigma_x^2}{n} = \frac{\mu_x}{n} \qquad (3)$$

Queste due relazioni definiscono l'errore standard e la varianza della media per popolazioni infinitamente grandi.

Esse mostrano che *la distribuzione delle medie campionarie diviene sempre più compatta attorno alla media della popolazione via via che la dimensione del campione n aumenta*, cioè la sua deviazione standard è inversamente proporzionale alla radice quadrata del numero di elementi nel campione.

Il grafico seguente mostra un confronto fra differenti distribuzioni di frequenza delle medie, ciascuna distribuzione essendo formata dalle medie di campioni aventi numerosità diversa:



Per esempio, supponiamo che la resa media di grano in chilogrammi per ettaro sia pari a 1000 con una deviazione standard di 50 Kg. Se consideriamo il peso medio della resa ottenuta in un campione di 25 campi di un ettaro ognuno, allora la deviazione standard per le medie di questi campioni sarà:

$$\sigma_m = \frac{\sigma_x}{\sqrt{n}} = \frac{50}{\sqrt{25}} = 10 \text{ kg.}.$$

Ma se, per contro, possiamo considerare campioni solo di 9 campi, la deviazione standard sarà:

$$\sigma_m = \frac{\sigma_x}{\sqrt{n}} = \frac{50}{\sqrt{9}} = 16,6 \text{ kg.}$$

2. Tecnica di stima nei grandi campioni

2.1 *Attendibilità della media nei g.c.*

Si supponga che da passate esperienze risulti che la media di un certo carattere di una popolazione di individui sia uguale a M, con una deviazione standard pari a σ.

Si supponga altresì che un campione di n individui della stessa popolazione sia stato assoggettato ad un particolare trattamento e che la media del carattere analizzato su essi dopo il trattamento sia risultata uguale a M_1. Si vuole conoscere fino a che punto questa differenza (M_1-M) nella misura del carattere relativa ad ogni individuo del campione può essere imputabile al nuovo trattamento praticato.

Poiché la media campionaria $m = M_1$ è distribuita normalmente, con media $M_m = M$ e deviazione standard $\sigma_m = \frac{\sigma_x}{\sqrt{n}}$, il valore dell'unità standard z corrispondente al valore di m per questo campione di n individui è:

$$z = \frac{m - M_m}{\sigma_m} = \frac{(M_1 - M)\sqrt{n}}{\sigma_x} = z_c .$$

Ora, dalle tavole che forniscono le aree sotto la curva normale, vediamo che la probabilità di avere per effetto del caso un valore di z maggiore di un limite prefissato z_0 è P_{z_0}, cioè si riscontra nel P_{z_0}% dei casi. In problemi analoghi a questo, generalmente si considera che una probabilità P_{z_0}, del 5% sia un limite significativo, che è chiamato "*limite di confidenza del 5%*". Quindi, se

$z_c > z_0$, dato che è molto piccola la probabilità di avere per effetto del caso un valore di z maggiore di z_c che risulta dal campione, si può ritenere che la differenza tra la media del campione e quella della popolazione sia significativa. Essa non può essere attribuita ad un errore di campionamento, perché solo P_{z_0}% volte su 100 di tali campioni si dovrebbe avere una media superiore ad M per effetto del caso. Quindi sembra verosimile ritenere che il nuovo trattamento a cui sono stati sottoposti quegli n individui abbia determinato un significativo mutamento nella media del carattere (aumento o diminuzione a seconda del segno della differenza.

Quando la deviazione standard della popolazione non è nota, se il campione è abbastanza grande, σ può essere sostituito da s, che è la deviazione standard calcolata sul campione.

Esempio VII.1

Da una esperienza del passato risulti che la resa settimanale media di latte in chilogrammi ottenuta da una mucca di 10 anni di età di una certa razza sia uguale a 80 kg, con una deviazione standard di 16 kg.

Si supponga altresì che un campione di 64 mucche della stessa età sia stato mantenuto sotto un nuovo regime alimentare e la resa settimanale media di latte ottenuta da ogni mucca del campione sia stata uguale a 85 Kg. Si vuole conoscere fino a che punto questo aumento settimanale di 5 Kg nella quantità di latte prodotto da ogni mucca può essere imputabile al nuovo regime alimentare adottato.

Con i dati noti si ha:

$$M = M_m = 80 \; ; \; m = 85 \; ; \; \sigma_m = \frac{16}{\sqrt{64}} = \frac{16}{8} = 2.$$

Poiché la media è distribuita normalmente, il valore dell'unità standard t corrispondente al valore di m per questo campione di 64 mucche è:

$$z = \frac{(m - M_m)}{\sigma_m} = \frac{(85 - 80)}{2} = 2{,}5.$$

Ora, dalle tavole che forniscono le aree sotto la curva normale, vediamo che la probabilità di avere per effetto del caso un valore di z maggiore di 2,5 è 0,0062, cioè é pari allo 0,62 percento. Quindi, dato che è molto piccola la probabilità di avere per effetto del caso un valore di z maggiore di 2,5 che risulta dal campione, si può ritenere che l'aumento nell'ammontare di latte prodotto settimanalmente sia significativo. Esso non può essere attribuito ad un errore di campionamento, perché solo 62 volte su 10000 di tali campioni si dovrebbe avere un peso medio del latte superiore ad 80 kg per effetto del caso. Quindi sembra verosimile

che il nuovo regime nutritivo a cui sono state sottoposte quelle 64 mucche abbia determinato un significativo incremento nella resa media in latte ottenuta da esse.

2.2 *Significatività della differenza tra due medie nei g.c.*

In molti casi pratici sorge il problema di sapere se, conoscendo la differenza tra le medie di due campioni, possiamo determinare la *significatività della differenza tra le medie delle popolazioni* da cui questi campioni sono stati tratti. Se indichiamo con x_{1i} ed x_{2i} le singole determinazioni dei due campioni, con n_1 ed n_2 le dimensioni dei due campioni, con m_1 ed m_2 le due medie, con s_1 ed s_2 le due deviazioni standard, mentre per le due popolazioni le singole determinazioni, le dimensioni rispettive, le medie e le deviazioni standard sono rappresentate da X_{1i} ed X_{2i}, N_1 ed N_2, M_1 ed M_2, da σ_1 e σ_2 ed inoltre se si indicano con M_{m1-m2} e σ_{m1-m2} la media e la deviazione standard della distribuzione delle differenze tra le due medie, allora è possibile enunciare il seguente teorema, la cui dimostrazione può essere trovata facilmente nella letteratura in materia:

Se m_1 ed m_2 sono distribuite normalmente ed indipendenti tra loro allora la distribuzione della differenza $(m_1 - m_2)$ è approssimativamente normale ed ha una media ed una deviazione standard uguali a:

$$M_{m_1-m_2} = M_1 - M_2 \qquad (4a)$$

$$\sigma_{m_1-m_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}\,. \qquad (4b)$$

Si supponga di dover ricercare se le medie delle due popolazioni sono uguali. Ci domandiamo, cioè, se è verificata la relazione $M_1 - M_2 = 0$ e per far ciò utilizziamo le conoscenze acquisite sulla base dei due campioni, quindi si usa la statistica $(m_1 - m_2)$ e si rigetta l'ipotesi nulla se questa differenza è significativamente diversa da zero. La distribuzione campionaria di $(m_1 - m_2)$ per campioni di grande dimensione è approssimativamente normale con media uguale a $(M_1 - M_2)$ e varianza uguale a

$$\sigma^2_{m_1-m_2} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

Allora, se $(M_1 - M_2) = 0$, possiamo supporre che la distribuzione campionaria della statistica:

$$z = \frac{m_1 - m_2}{\sigma_{m_1-m_2}} = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \qquad (5)$$

sia approssimativamente rilevabile dalle tavole dei valori cumulativi della distribuzione normale ossia dalle tavole che danno il valore dell'area sottostante la curva normale a partire da - ∞ e fino ad un determinato valore di z.

I casi che si possono presentare nella pratica possono essere ben diversi tra loro quanto alla qualità dei dati noti a disposizione e per ciascun caso, allora, si dovrà seguire un procedimento ben definito, come verrà detto qui di seguito.

I caso. *Le medie di due popolazioni sono uguali ed hanno la stessa varianza σ^2, che si suppone nota.*

In questo caso, dunque, si ha la seguente situazione.

a) L'ipotesi H che si fa è: $M_1 = M_2 = M$ ed anche $\sigma^2_1 = \sigma^2_2 = \sigma^2$, cioè si suppone l'uguaglianza delle medie tra di loro e con la media generale e delle varianze tra di loro e con la varianza generale.

b) Si prende una decisione sull'errore del I tipo α, il quale, come si è detto più volte, misura la probabilità di rigettare l'ipotesi quando essa è vera (vedasi il capitolo VI).

c) Si usa la statistica z descritta dalla (5) con le opportune semplificazioni.

d) Se l'universo è distribuito normalmente, la distribuzione campionaria di z è normale; oppure, in parecchi casi, se le dimensioni dei campioni N_1 ed N_2 sono grandi, la distribuzione campionaria di z è approssimativamente normale.

e) La regione critica è definita dalla relazione:

$$z_{1-\frac{1}{2}\alpha} < z < z_{\frac{1}{2}\alpha}$$

Come applicazione di questo caso si consideri il seguente esempio.

Esempio VII.2

Due astronomi hanno raccolto osservazioni su una data stella. Le 12 osservazioni raccolte dal primo astronomo hanno una misura media di 1,20. Le 8 osservazioni ottenute dal secondo astronomo hanno una media di 1,15. L'esperienza passata ha indicato che questi astronomi ottengono misure con una varianza di circa 0,40. Ci si chiede se la differenza tra i due risultati è significativa.

a) H: $M_1 = M_2$, $\sigma^2 = 0{,}40$, $\sigma = \sqrt{0{,}40}$,= 0,63245.

b) Si sceglie $\alpha = 0{,}01$.

c) Si utilizza la statistica:

$$z = \frac{m_1 - m_2}{\sigma}\left[\frac{1}{N_1} + \frac{1}{N_2}\right]^{-\frac{1}{2}}$$

d) Si suppone che z abbia una distribuzione normale con media 0 e varianza 1.

e) Si rigetta l'ipotesi H se $z < -2{,}58$ oppure se $z > 2{,}58$.

f) Si calcola z con la formula sub c) ottenendo:

$$z = \frac{1{,}20 - 1{,}15}{0{,}6325\sqrt{\frac{1}{12} + \frac{1}{8}}} = 0{,}17$$

che è compreso tra - 2,58 e + 2,58 e quindi accettiamo l'ipotesi. Dunque, la differenza tra le due medie non è significativa.

II caso. *La media di una popolazione è minore o uguale alla media dell'altra popolazione e la varianza σ^2 è nota*.

Spesso si incontrano problemi che riguardano la ineguaglianza tra le medie di due popolazioni. In questi casi, la formulazione dell'ipotesi viene fatta nella forma H: $M_1 \leq M_2$, in cui si considera che l'uguaglianza è in dubbio e che il segno di minore o uguale ci deve ricordare che si rigetterà l'ipotesi se e solo se m_1 è significativamente più grande di m_2. I passi da compiere sono i seguenti.

a) H: $M_1 \leq M_2$, data l'eguaglianza $\sigma^2_1 = \sigma^2_2 = \sigma^2$ e tenuto conto che il valore di σ^2 è noto.

b) Si sceglie α.

c) La statistica da usare in questo caso è ancora:

$$z = \frac{m_1 - m_2}{\sigma}\left[\frac{1}{N_1} + \frac{1}{N_2}\right]^{-\frac{1}{2}}$$

d) Si suppone che questa statistica abbia una distribuzione normale con media zero e varianza 1.

e) Non si vuole rigettare l'ipotesi se m_1 risulta minore di m_2 e così si assume come regione critica l'intervallo $z < z_{1-\alpha}$

f) Si calcola la statistica e si rigetta o si accetta l'ipotesi a seconda del risultato.

III caso. *Due popolazioni hanno la stessa media, e la stessa varianza, ma σ^2 non è noto*.

I passi successivi da compiere sono i seguenti.

a) H: $M_1 = M_2$. Si sa che $\sigma^2_1 = \sigma^2_2 = \sigma^2$, ma non il valore di σ^2.

b) Si sceglie α.

c) Come statistica per verificare questa ipotesi si usa:

$$t = \frac{m_1 - m_2}{s_p}\left[\frac{1}{N_1} + \frac{1}{N_2}\right]^{-\frac{1}{2}}$$

in cui s^2_p è la stima di σ^2 definita dalla relazione:

$$s_p^2 = \frac{\sum X_{1i}^2 - \frac{\left(\sum X_{1i}\right)^2}{N_1} + \sum X_{2i}^2 - \frac{\left(\sum X_{2i}\right)^2}{N_2}}{N_1 + N_2 - 2} = \frac{(N_1 - 1).s_1^2 + (N_2 - 1).s_2^2}{N_1 + N_2 - 2}$$

Il primo sommatorio che figura nella prima frazione rappresenta la somma dei quadrati relativi al primo campione, il terzo sommatorio l'analoga quantità relativa al secondo campione, il secondo sommatorio (tra parentesi tonde) rappresenta la somma delle osservazioni nel primo campione ed il quarto (sempre tra parentesi tonde) rappresenta la somma delle osservazioni relative al secondo campione.

d) Se ambedue le popolazioni seguono distribuzioni normali con la stessa media e la stessa varianza, allora la statistica t segue una distribuzione con un numero di gradi di libertà pari a $(N_1 + N_2 - 2)$.

e) La regione di rigetto è definita dalla relazione:

$$ {}_{(N_1+N_2-2)}t_{1-\frac{1}{2}\alpha} < t < {}_{(N_1+N_2-2)}t_{\frac{1}{2}\alpha} $$

f) Si calcola t e si rigetta o si accetta l'ipotesi, a seconda del risultato.

Come applicazione si consideri il seguente esempio.

Esempio VII.3

Due differenti tipi di razioni alimentari vengono somministrate ai maiali di una porcilaia. Si vuole verificare quale dei due tipi sia migliore. Un campione di 12 maiali viene alimentato con la razione di tipo A ed un altro campione di 12 maiali riceve, invece, la razione di tipo B. I guadagni in peso registrati sono i seguenti:

Tipo A	Tipo B
31	26
34	24
29	28
26	29
32	30
35	29
38	32
34	26
30	31
29	29
32	32
31	28

a) H: $M_1 = M_2$.
b) Si sceglie $\alpha = 0{,}05$.
c) Si utilizza la statistica:

$$ t = \frac{m_1 - m_2}{s_p}\left[\frac{1}{N_1} + \frac{1}{N_2}\right]^{-\frac{1}{2}} $$

d) Se le due popolazioni hanno distribuzioni normali con la stessa media e la stessa varianza, allora questa statistica ha una distribuzione t con 12+12-2 = 22 gradi di libertà.
e) Si rigetta l'ipotesi H se $t < -2{,}07$ oppure se $t > 2{,}07$.

Per il tipo A la media sarà $m_1 = 31{,}7500$ mentre per la somma dei quadrati si avrà:

$$ (N_1 - 1).s_1^2 = \sum X_{1i}^2 - \frac{\left(\sum X_{1i}\right)^2}{N_1} = 112{,}25 $$

Per il tipo B la media sarà $m_2 = 28{,}6667$ mentre per la somma dei quadrati si avrà:

$$ (N_2 - 1).s_2^2 = \sum X_{2i}^2 - \frac{\left(\sum X_{2i}\right)^2}{N_2} = 66{,}64. $$

Per cui, sulla base dei risultati suddetti si può calcolare ora:

$$ s_p^2 = \frac{(N_1 - 1).s_1^2 + (N_2 - 1).s_2^2}{N_1 + N_2 - 2} = \frac{112{,}25 + 66{,}64}{12 + 12 - 2} = 8{,}131 $$

e quindi

$$ s_p = \sqrt{8{,}131} = 2{,}85. $$

f) Si calcola t con la formula sub c) ottenendo:

$$ t = \frac{31{,}7500 - 28{,}6667}{2{,}85.\sqrt{\frac{1}{12} + \frac{1}{12}}} = \frac{3{,}0800.\sqrt{6}}{2{,}85} = 2{,}65 $$

che è un valore maggiore di 2,07 e quindi si deve rigettare l'ipotesi di parità dei guadagni in peso e si stabilisce che la razione di tipo A tende a generare guadagni in peso superiori a quelli che può determlinare l'uso della razione di tipo B.

IV caso. *La media di una popolazione è minore o uguale alla media di una seconda popolazione e la varianza σ^2 è ignota*.

Questo problema è una combinazione di quelli considerati negli esempi che precedono. In questo caso si userà lo stesso test statistico utilizzato in precedenza, ma si dovrà rigettare l'ipotesi solo se è:

$$ {}_{(N_1+N_2-2)}t_{1-\alpha} < t $$

Come applicazione consideriamo il seguente esempio.

Esempio VII.4

Due tipi di vernici devono essere confrontati. Il tipo I è un poco più a buon mercato del tipo II. Il test consiste nell'assegnare un punteggio alle due vernici, dopo che esse sono state esposte a certe condizioni atmosferiche per un periodo di 6 mesi. Di ciascun tipo di vernice vengono esaminati 5 campioni ai quali risultano attribuiti i punteggi seguenti:

Tipo I	Tipo II
85	89
87	89
92	90
80	84
84	88

In linea di principio la preferenza viene data al tipo I che è il meno costoso, a meno che non vi sia una ragione precisa per ritenere che il tipo II sia migliore. Se si ammette l'ipotesi che sia $M_2 \leq M_1$, la probabilità di adottare il tipo II quando le vernici sono ugualmente buone può essere controllata al livello α.

a) H: $M_2 \leq M_1$.

b) Livello di $\alpha = 0,05$.

c) Si utilizza la statistica:

$$t = \frac{m_1 - m_2}{s_p}\left[\frac{1}{N_1} + \frac{1}{N_2}\right]^{-\frac{1}{2}}$$

d) Si suppone che questa statistica abbia una distribuzione come la t con $(N_1 + N_2 - 2)$ gradi di libertà.

e) Si rigetta l'ipotesi H se $t < t_{0,05}$ e cioè se $t < -1,86$.In questo caso si rigetta l'ipotesi se t è minore di - 1,86, perchè questo fatto induce a credere che sia $M_2 > M_1$.

f) Si calcola t con la formula sub c) tenendo presente che i calcoli intermedi forniscono i risultati seguenti: per le medie si ha $m_1 = 85,6$ ed $m_2 = 88,0$; per la somma dei quadrati del tipo I si ha invece:

$$(N_1 - 1)s_1^2 = \sum X_{1i}^2 - \frac{(\sum X_{1i})^2}{N_1} = 77,2$$

e per il tipo II :

$$(N_2 - 1)s_2^2 = \sum X_{2i}^2 - \frac{(\sum X_{2i})^2}{N_2} = 22,0.$$

Per cui, sulla base dei risultati suddetti si può calcolare ora:

$$s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} = \frac{77,2 + 22,0}{8} = 12,4$$

e quindi

$$s_p = \sqrt{12,4} = 3,52.$$

$$t = \frac{85,6 - 88,0}{3,52.\sqrt{\frac{1}{5} + \frac{1}{5}}} = \frac{-2,4.\sqrt{5}}{3,52} = -1,08$$

che è un valore maggiore di - 1,86 e quindi si deve accettare l'ipotesi e concludere che poiché la vernice di tipo II pur essendo più costosa non assicura risultati migliori, è meglio utilizzare quella di tipo I che assicura risultati analoghi, ma costa meno.

V caso. *Due popolazioni hanno la stessa media, ma varianze diverse*.

Si supponga di voler comparare le medie di due popolazioni distribuite normalmente che abbiano varianze pari a σ^2_1 e, rispettivamente, σ^2_2. Come precedentemente, indichiamo le osservazioni con X_{1i} ed X_{2i} e le dimensioni dei campioni con N_1 ed N_2. La distribuzione campionaria della statistica:

$$z = \frac{(m_1 - m_2) - (M_1 - M_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

è normale con media 0 e varianza 1. Questo è approssimativamente vero per valori sufficientemente grandi di N_1 ed N_2, anche se le popolazioni non sono normali.

Se i valori di σ_1 e σ_2 sono noti, essi sono sostituiti nella formula ed i valori osservati di z possono essere comparati con i valori cumulativi della distribuzione normale.

Se i valori di σ_1 e σ_2 non sono noti, ma chi esegue l'esperimento ritiene che vi siano elementi sufficientemente evidenti per ammettere che ognuna delle due popolazioni sia normalmente distribuita, allora è possibile usare la statistica ottenuta sostituendo i valori osservati di s_1 ed s_2 al posto di σ_1 e σ_2, pervenendo alla statistica:

$$t = \frac{(m_1 - m_2) - (M_1 - M_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

che, se le ipotesi di normalità sono corrette, ha approssimativamente una distribuzione t con v gradi di libertà, dove é:

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}} - 2.$$

In generale, questo valore di v non sarà un intero e quindi potrà essere necessaria l'interpolazione tra i valori contenuti nella tavola di t per dati livelli di significatività, usando il metodo di interpolazione per punti noti ed adattando una funzione lineare in due variabili.

Tuttavia, non sempre è necessario procedere al calcolo suddetto ed anzi, di solito, è sufficiente utilizzare il valore della tavola più vicino a quello calcolato.

Esempio VII.5

Si abbiano le seguenti indicazioni relative alle rese del grano in due provincie. Esistono elementi per sospettare che le rese medie siano uguali, cioè $M_1 = M_2$ e quindi $M_{m1 - m2} = 0$.

Si supponga che in ciascuna provincia si sia formato un campione di particelle a grano in numero di $n_1 = 100$ nella prima provincia ed $n_2 = 200$ nella seconda. Le rese medie dei due campioni siano $m_1 = 500$ ed $m_2 = 520$, mentre le rispettive deviazioni standard siano $s_1 = 80$ ed $s_2 = 120$.

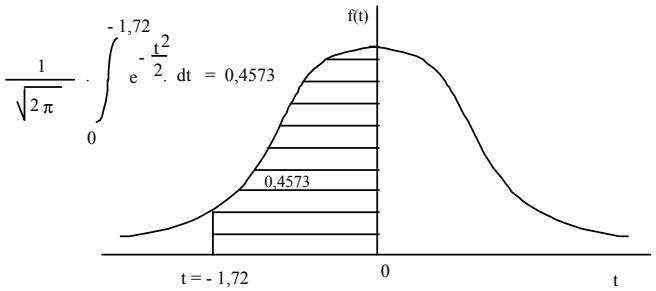
Poiché σ_1 e σ_2 non sono noti e dato che la dimensione del campione è sufficientemente grande, è possibile sostituire alle varianze delle popolazioni quelle dei rispettivi campioni. Si avrà quindi:

$$s_{m_1 - m_2} = \sqrt{\frac{80^2}{100} + \frac{120^2}{200}} = 11{,}6$$

Il corrispondente valore di t è il seguente:

$$t = \frac{(m_1 - m_2) - (M_1 - M_2)}{s_{m_1 - m_2}} = \frac{500 - 520}{11{,}6} = -1{,}72.$$

Ora, nella tavola delle aree della curva normale la probabilità corrispondente a $t = -1{,}72$ è:



Da questo risultato è possibile trarre le seguenti conclusioni: la probabilità che la differenza tra le medie di qualunque coppia di campioni sia maggiore della differenza ottenuta dai dati considerati, cioè $|m_1 - m_2| = 20$, è uguale a:

$$1 - 2 \times 0{,}4573 = 1 - 0{,}9146 = 0{,}0854,$$

ossia l'area sotto la curva normale che ricade al di fuori dell'intervallo (-1,72, +1,72).

Poiché questa probabilità 0,0854 è più grande di quella relativa all'intervallo di confidenza calcolato a livello del 5%, le differenze tra le rese delle due provincie saranno considerate non significative al livello del 5% di significatività.

In altre parole, la ipotesi $M_1 - M_2 = 0$ è giustificata.

2.3 *Accoppiamento delle osservazioni*

Se è evidente che σ_1 e σ_2 non sono uguali, a volte risulta più conveniente adottare un piano di campionamento ed eseguire le analisi necessarie procedendo all'accoppiamento delle osservazioni tratte dalle due popolazioni. Così pure, qualche volta, nell'estrarre campioni da due popolazioni capita che fattori esterni causino una differenza significativa nelle medie, mentre non vi è alcuna differenza negli effetti che si vogliono misurare.

Per esempio, se si vuol conoscere tra due diversi metodi di insegnamento quale sia migliore, si possono prendere due gruppi di studenti ed insegnare ad un gruppo con il primo metodo ed all'altro gruppo con il secondo metodo. Se uno dei gruppi è formato da studenti migliori (o più maturi o meglio addestrati nelle materie di base, eccetera) di quelli dell'altro gruppo, i risultati dell'esperimento possono non riflettere la efficacia dei metodi di insegnamento.

Un altro esempio è il seguente.

Supponiamo che in un esperimento per verificare quale tra due tipi (A o B) di fertilizzanti sia migliore, in 10 stazioni sperimentali siano stati seminati a grano due

appezzamenti di terreno per ciascuno. Su uno dei due appezzamenti sia stato sparso il fertilizzante A e sull'altro il fertilizzante B. Se le medie dei 10 appezzamenti fertilizzati con il tipo A fossero confrontate con le medie dei 10 appezzamenti che hanno ricevuto il tipo B, parte della differenza osservata (se ve ne fosse una) potrebbe essere dovuta al diverso tipo di suolo o a diverse condizioni meteorologiche invece che ai due diversi fertilizzanti. Un'altra possibilità è quella che i fertilizzanti causino una differenza, ma questa differenza risulti nascosta da altri fattori.

Un piano degli esperimenti che talvolta permette di superare parte delle suddette difficoltà è quello che si basa sull'espediente di accoppiare le osservazioni, facendo in modo che le due componenti di ciascuna coppia siano simili sotto tutti gli aspetti tranne quello che si sta esaminando. Questa situazione sarebbe l'ideale, naturalmente, ma nella generalità dei casi si è limitati dalla disponibilità di coppie che siano simili e dalla capacità che chi effettua l'esperimento ha di scegliere coppie simili.

Applicando questo criterio, nel primo esempio dovrebbero essere prese coppie di studenti di abilità approssimativamente uguale e, mentre uno dei componenti di ciascuna coppia dovrebbe essere assoggettato al primo metodo, l'altro dovrebbe essere sottoposto al secondo metodo. Nel secondo esempio, ogni coppia di appezzamenti dovrebbe avere approssimativamente lo stesso tipo di suolo, essere esposto a condizioni meteorologiche identiche, eccetera.

Si supponga di indicare con il simbolo X_{i1} il primo componente della i-esima coppia e con X_{i2} il secondo componente. Abbiamo N coppie di osservazioni:

$$(X_{11}, X_{12}),\ (X_{21}, X_{22}), (X_{31}, X_{32}), ..., (X_{N1}, X_{N2}),.$$

Se si considerano le differenze $(X_{i1} - X_{i2}) = d_i$, si avrà un insieme di N osservazioni, ciascuna delle quali è una differenza tra due osservazioni originarie. Potranno esservi dei fattori esterni che influiscono su alcuni dei soggetti individuali esaminati, ma si suppone che essi influiscano su ogni componente di qualunque coppia esattamente nello stesso modo. Inoltre si suppone anche che l'effetto sia essenzialmente quello di aumentare (o diminuire) ognuna delle medie di un fattore costante in modo che la differenza tra le medie annulli l'effetto.

Si voglia verificare l'ipotesi che sia $M_1 = M_2$. Questa ipotesi afferma che non vi è differenza tra trattamenti e cioè che non vi è differenza all'interno delle coppie. Vi è probabilmente una differenza tra coppia e coppia, ma si suppone che questa all'interno di ogni singola coppia sia stata eliminata tramite l'uso della differenza $(X_{i1} - X_{i2})$. Per verificare la validità della suddetta ipotesi, si noti che, se essa fosse vera, le differenze $(X_{i1} - X_{i2})$ dovrebbero provenire da un insieme di numeri la cui media è 0. Quindi è possibile verificare l'ipotesi che $(X_{i1} - X_{i2})$ provenga da un universo con media $M = 0$, utilizzando la statistica t con $(N-1)$ gradi di libertà, ove N é il numero delle coppie considerate.

Un altro vantaggio di questo metodo è che non dobbiamo supporre che le due varianze σ^2_1 e σ^2_2 siano uguali o che i valori X_{i1} e X_{i2} siano indipendenti. Nel paragrafo precedente, invece, si era ipotizzato che le osservazioni fossero indipendenti, cioè che ogni singola osservazione corrispondeva ad un singolo elemento selezionato casualmente.

Se non vi sono effetti esterni apprezzabili, in realtà quando si procede all'accoppiamento delle osservazioni si perdono informazioni. Questa perdita si traduce in un aumento della probabilità di accettare l'ipotesi quando essa è falsa (errore del II tipo). L'aumento è lieve, tuttavia, se le dimensioni del campione sono moderatamente grandi, ad esempio maggiori di 10, e il livello di significatività non ne risulta influenzato.

Con soli $(N-1)$ gradi di libertà nella stima di σ^2 si accettano differenze in $(m_1 - m_2)$ più grandi di quelle che si potrebbero accettare se si avessero $(2N-2)$ gradi di libertà. Naturalmente, se vi sono effetti estranei (come negli esempi fatti in precedenza) si potrebbe essere obbligati ad accoppiare le osservazioni ed a correre il rischio di una perdita di precisione.

Esempio VII.6

In uno studio sulla capacità di apprendimento, da una classe di matricole sono stati scelti a caso 10 ragazzi e 10 ragazze. Ad essi fu attribuito un punteggio, misurando la loro abilità ad apprendere sillabe senza senso. Poiché il conduttore dell'esperimento sospettava che la varianza sarebbe stata diversa per ragazzi e ragazze (σ^2_1 diverso da σ^2_2), prese la decisione di accoppiare i dati.

L'analisi che qui viene mostrata è appropriata per un accoppiamento casuale dei soggetti. Tuttavia, è possibile che l'accoppiamento dei soggetti sulla base del quoziente di intelligenza, per esempio, determini una riduzione nella varianza della popolazione e quindi renda più facile scoprire qualunque differenza in abilità. Le osservazioni e le analisi sono riportate qui di seguito.

	Numero d'ordine della coppia									
	1	2	3	4	5	6	7	8	9	10
Ragazzi	28	18	22	27	25	30	21	21	20	27
Ragazze	19	38	42	25	15	31	22	37	30	24
Differenza	9	- 20	- 20	2	10	- 1	- 1	- 16	- 10	3

$\Sigma d_i = -44$; $\Sigma d^2_i = 1352$; $m_d = -4,4$; $s^2_d = (1352 - 44^2/10)/9 = 128,7$.

a) H: $M_1 = M_2$, cioè il punteggio medio di apprendimento per i ragazzi è uguale a quello delle ragazze.

b) Si sceglie $\alpha = 0,01$.

c) Si utilizza la statistica:

$$t = \frac{(m_d - 0)}{s / \sqrt{N}} = \frac{-4,4}{\sqrt{128,7} / \sqrt{10}} = -1,2.$$

(Si noti che è $m_1 - m_2 = m_d$ e che $s_d{}^2$ è la varianza delle differenze).

d) Se le due popolazioni hanno distribuzioni normali, allora questa statistica ha una distribuzione t con 10-1 = 9 gradi di libertà.

e) Si rigetta l'ipotesi H se $t < -3,25$ oppure se $t > 3,25$.

f) Si calcola t con la formula sub c) ottenendo $t = -1,2$ che è un valore compreso nell'intervallo indicato e quindi si deve accettare l'ipotesi.

L'esperimento condotto, pertanto, fornisce come risultato che non esiste una differente abilità ad imparare sillabe senza senso relativamente ai ragazzi ed alle ragazze considerate nel campione.

Esempio VII.7

Un certo stimolo deve essere verificato per i suoi effetti sulla pressione del sangue. Su 12 uomini viene misurata la pressione sanguigna prima e dopo lo stimolo. I risultati sono i seguenti.

Soggetti esaminati	Prima	Dopo	Differenza d_i	d^2_i
1	120	128	+8	64
2	124	131	+7	49
3	130	131	+1	1
4	118	127	+9	81
5	140	132	-8	64
6	128	125	-3	9
7	140	141	+1	1
8	135	137	+2	4
9	125	118	-8	64
10	130	132	+2	4
11	126	129	+3	9
12	127	135	+8	64

22 414

Vi è motivo di credere che lo stimolo elevi la pressione sanguigna in media di 5 punti?

L'accoppiamento qui si rivela necessario poichè sullo stesso individuo sono effettuate due osservazioni. Il campione consiste di 12 individui con due misurazioni ciascuno.

a) H: $M_2 - M_1 \leq 5$ (cioè la media delle differenze è uguale a 5).
b) Si sceglie $\alpha = 0{,}05$.
c) Si utilizza la statistica:

$$t = \frac{(m_1 - m_2) - 5}{s / \sqrt{12}} = \frac{m_d - 5}{s / \sqrt{12}}$$

[Si noti che è $(m_1 - m_2) = m_d$ e che s_d è la deviazione standard delle differenze].

d) Se le due popolazioni hanno distribuzioni normali e se gli effetti esterni sono additivi, allora questa statistica ha una distribuzione t con (12 - 1) = 11 gradi di libertà.
e) Si rigetta l'ipotesi H se $t > +1{,}80$.
f) Si calcola $s_d^2 = [414 - (22)^2/12]/11 = 33{,}97$ e quindi $s_d = 5{,}83$. Poichè $m_d = 22/12 = 1{,}833$, allora con la formula sub c) si otterrà $t = -1{,}9$ e quindi si deve accettare l'ipotesi, cioè che lo stimolo ha per effetto quello di elevare la pressione in media di 5 punti.

2.4 *Significatività della differenza tra percentuali nei g. c.*

In alcuni problemi spesso si devono usare misure percentuali, invece di valori assoluti. Se indichiamo con p'_1 e p'_2 le percentuali relative a due campioni basate su n_1 ed n_2 prove e se p_1 e p_2 rappresentano le percentuali delle corrispondenti popolazioni, allora la media e la deviazione standard delle differenze tra percentuali sono date dalle seguenti relazioni:

$$M_{p'_1 - p'_2} = p_1 - p_2 \qquad \text{e} \qquad \sigma_{p'_1 - p'_2} = \sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}$$

in cui $q_1 = (1 - p'_1)$ e $q_2 = (1 - p'_2)$.

Esempio VII.8

Per mostrare l'applicazione della formula precedente, si consideri l'esempio che segue. Due diversi tipi di siero sono somministrati a 134 mucche colpite da una malattia. Il siero A è iniettato a 50 mucche e 5 di queste muoiono. Il siero B è iniettato alle restanti 84 mucche e 6 di esse muoiono. Si vuol sapere se la differenza di effetto tra i due tipi di siero sulla mortalità delle mucche per la malattia in questione è significativa.

Il tasso di mortalità corrispondente al siero del tipo A è 5/50 = 1/10 ossia il 10% ovvero $p'_1 = 0{,}10$. Invece, il tasso di mortalità corrispondente al siero del tipo B è 6/84 = 1/14 = 0,0714 ossia il 7,14% ovvero $p'_2 = 0{,}0714$. Il numero totale di mucche malate è 134, mentre complessivamente le mucche morte sono 11. Quindi, avremo $p = 11/134 = 0{,}082$ cioè 8,2% e $q = 1 - 0{,}082 = 0{,}918$. In questo problema p_1 e p_2 non sono noti.

Se si suppone che $p_1 = p_2 = p$, sarà anche $q_1 = q_2 = q$ e si avrà:

$$M_{p'_1 - p'_2} = p_1 - p_2 = 0$$

$$\sigma_{p'_1 - p'_2} = \sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}} = \sqrt{0{,}082 \text{x} 0{,}918 \text{x} \left(\frac{1}{50} + \frac{1}{84}\right)}$$

cioè

$$s_{p'_1 - p'_2} = \sqrt{\frac{0{,}082 \text{x} 0{,}918 \text{x} 134}{4200}} = \sqrt{0{,}002401663} = 0{,}049.$$

Con questi risultati, si ottiene, quindi:

$$t = \frac{0{,}10 - 0{,}0714}{0{,}049} = \frac{0{,}0286}{0{,}049} = 0{,}58367$$

in corrispondenza del quale si trova:

$$\frac{1}{\sqrt{2\pi}} \int_0^{0{,}58367} e^{-\frac{t^2}{2}} \cdot dt = 0{,}2202.$$

Quindi la probabilità di avere una differenza maggiore della differenza (0,10 - 0,07) = 0,03 è pari a (1 - 2 . 0,2202) = (1 - 0,4404) = 0,5596. Si conclude, dunque, che circa 56 coppie di campioni casuali su 100 dovrebbero dare una differenza più grande di 0,03, cosicché la differenza tra i due sieri non è affatto significativa.

2.5 *Distribuzione della deviazione standard nei g. c.*

La distribuzione di frequenza delle deviazioni standard di tutti i possibili campioni estratti da una popolazione avrà una media M_s approssimativamente uguale alla deviazione standard σ della popolazione se la dimensione dei campioni è sufficientemente grande. D'altra parte, si può dimostrare che la deviazione standard di questa distribuzione è data da σ_s e dunque le due statistiche saranno:

$$M_s = \sigma \text{ e} \qquad \sigma_s = \frac{\sigma}{\sqrt{2n}}$$

Inoltre, per grandi campioni la distribuzione della deviazione standard diviene approssimativamente normale, quindi, di tutte le deviazioni standard campionarie, il 95% cadrà nell'intervallo:

$$\sigma - 2.\sigma_s \le M_s = \sigma \le \sigma + 2.\sigma_s$$

cioè:

$$\sigma - 2.\frac{\sigma}{\sqrt{2n}} \le M_s = \sigma \le \sigma + 2.\frac{\sigma}{\sqrt{2n}}$$

ed il 99% cadrà nell'intervallo:

$$\sigma - 3.\frac{\sigma}{\sqrt{2n}} \le M_s = \sigma \le \sigma + 3.\frac{\sigma}{\sqrt{2n}}$$

Se σ non è noto e se il campione è sufficientemente grande, la formula scritta per σ_s può essere scritta come segue:

$$\sigma_s = \frac{s}{\sqrt{2n}}$$

Per esempio, se la deviazione standard di un campione di 200 elementi risulta uguale a $s = 6,5$ si ha:

$$\sigma_s = \frac{6,5}{\sqrt{2x200}} = \frac{6,5}{\sqrt{400}} = 0,325.$$

Usando l'intervallo di confidenza del 95% è possibile dire che la deviazione standard della popolazione sarà compresa tra i limiti:

$$6,5 - 2x0,325 \le \sigma \le 6,5 + 2x0,325$$

$$5,85 \le \sigma \le 7,15$$

2.6 *Differenza tra due deviazioni standard nei g. c.*

L'errore standard della differenza tra due deviazioni standard campionarie indipendenti è dato dalla formula seguente:

$$\sigma_{s_1 - s_2} = \sqrt{\sigma_{s_1}^2 + \sigma_{s_2}^2}$$

Si supponga di avere due grandi campioni di numerosità rispettive $n_1 \ge 30$ (o 50) ed $n_2 \ge 30$ (o 50). Le rispettive medie e deviazioni standard siano m_1, m_2, s_1 ed s_2. Supponiamo inoltre che per la popolazione sia $\sigma_1 = \sigma_2$. In base ai dati noti, si avrà:

$$t = \frac{(s_1 - s_2) - M_{s_1 - s_2}}{\sigma_{s_1 - s_2}} = \frac{(s_1 - s_2) - 0}{\sqrt{\sigma_{s_1}^2 + \sigma_{s_2}^2}} = \frac{(s_1 - s_2)}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}}$$

Ora, dalla tavola della curva normale standardizzata si ha che per il valore di t fornito dalla formula suddetta l'integrale tra zero e t ha un certo valore che indica la probabilità di avere un valore di t inferiore a quello calcolato. Se il complemento all'unità di questa probabilità risulta più piccolo del livello di confidenza prefissato, allora ciò vuol dire che è difficile ottenere per effetto del caso un valore di t superiore a quello calcolato e quindi si può concludere che l'ipotesi di uguaglianza tra le deviazioni standard delle due popolazioni non è giustificata, perchè tra esse esiste una differenza significativa. Se, invece, tale complemento all'unità risulta più grande del livello di confidenza prefissato, non si può escludere che effettivamente le due deviazioni standard siano uguali.

Esempio VII.9

Per illustrare anche questo caso si consideri il seguente esempio. Si supponga di avere due grandi campioni per i quali le numerosità, le medie e le deviazioni standard risultino essere le seguenti: $n_1 = 329$; $n_2 = 302$; $m_1 = 32,75$; $m_2 = 30,6$; $s_1 = 8,05$; $s_2 = 6,95$. E si supponga inoltre che sia $\sigma_1 = \sigma_2$. Con queste posizioni applicando le formule indicate, si avrà:

$$t = \frac{(s_1 - s_2)}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{8,05 - 6,95}{\sqrt{\frac{8,05^2}{2x329} + \frac{6,95^2}{2x302}}} = \frac{1,10}{0,422439} = 2,6039.$$

Ora, dalla tavola della curva normale standardizzata si ha che per $t = 2,6$ l'integrale tra zero e t vale:

$$\frac{1}{\sqrt{2\pi}} \int_0^{2,6} e^{-\frac{t^2}{2}} . dt = 0,4953$$

e quindi:

$$1 - \frac{1}{\sqrt{2\pi}} \int_{-2,6}^{2,6} e^{-\frac{t^2}{2}} . dt = 1 - 2 x 0,4953 = 0,0094.$$

Questa probabilità è più piccola del 5% ed infatti è 0,0094 < 0,05, per cui l'ipotesi che sia $\sigma_1 = \sigma_2$ non è giustificata. In altre parole, come già detto, esiste una differenza significativa tra le deviazioni standard delle due popolazioni.

3. Tecnica di stima nei piccoli campioni

3.1 *Attendibilità della media nei p. c. (t di Student)*

Si è visto prima che *la media di un campione è una stima non distorta della media della popolazione,* ma che *la deviazione standard è distorta nel senso che sottostima il valore della deviazione standard relativa alla popolazione*. Questa distorsione aumenta via via che diminuisce la dimensione del campione.

Teoricamente, si può mostrare che il valore atteso (o speranza matematica o valore medio) della varianza di un campione di n elementi è data dalla relazione seguente:

$$E\left(s^2\right) = \frac{n-1}{n}\sigma^2$$

in cui il fattore $(n-1)/n$ è chiamato "*correzione di Bessel*".

Di converso, la migliore stima della varianza della popolazione è definita come segue:

$$\hat{\sigma}^2 = \frac{n}{n-1}s^2 = \frac{\sum_{i=1}^{i=n}\left(x_i - m_x\right)^2 . f_i}{n-1}$$

e quindi la deviazione standard sarà:

$$\hat{\sigma} = s\sqrt{\frac{n}{n-1}} = \sqrt{\frac{\sum_{i=1}^{i=n}\left(x_i - m_x\right)^2 . f_i}{n-1}} .$$

Quando si abbia a che fare con campioni di dimensione minore di 50 elementi (in certi casi anche di 30 elementi), se la varianza della popolazione non è nota, essa può essere sostituita dalla migliore stima del suo valore indicata col segno $\hat{\sigma}^2$.

Per esempio, la formula che definisce la deviazione standard delle medie dei campioni nel caso di piccoli campioni assumerà la forma seguente:

$$\sigma_m = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n-1}} .$$

Nel caso dei piccoli campioni, la curva normale non fornisce i migliori risultati per determinare il grado di attendibilità della media campionaria o la significatività della differenza tra due medie. Per piccoli valori di n si usa un'altra distribuzione campionaria che considera la deviazione standard del campione piuttosto che quella della popolazione.

Questa nuova distribuzione, che non è strettamente normale, come si è già visto, è chiamata "*distribuzione t di Student*" o più semplicemente "*distribuzione della t*" e dipende da un parametro v che rappresenta i gradi di libertà (si veda capitolo IV, paragrafo 15).

L'uso della distribuzione della t in problemi che sfruttano piccoli campioni è illustrato nell'esempio che segue.

Esempio VII.10

Dall'esperienza passata si sa che il numero medio di petali in un certo tipo di fiore è pari a 30. Un campione di 10 fiori viene sottoposto ad un trattamento molto costoso ed il numero medio di petali in questo campione risulta essere pari a 33, con una deviazione standard di 2. L'aumento del numero dei petali è significativo?

Si sa che in questo problema n = 10, M = 30, m = 33 ed s = 2. Applicando la formula da adottare nel caso dei piccoli campioni, abbiamo:

$$t = \frac{m - M}{\frac{s}{\sqrt{n-1}}} = \frac{(33-30)x\sqrt{9}}{2} = 4,5$$

In questa formula è $v = n - 1 = 10 - 1 = 9$. Quindi cercando sulla tavola il valore corrispondente a $t = 4,5$ e $v = 9$ si trova che la probabilità di una deviazione maggiore di $t = 4,5$ è minore di 0,005. Questo valore è più piccolo del livello di significatività del 5% ed anche dell' 1%.

Il risultato, quindi, ci dice che l'aumento nel numero di petali è molto significativo e può essere attribuito all'effetto del trattamento costoso.

3.2 *Limiti di confidenza della media nei p. c.*

Per un campione di $n < 30$ (o 50) elementi si abbiano la media m, la sua deviazione standard s ed un numero di gradi di libertà pari a $v = (n-1)$.

L'intervallo di confidenza della media ad un livello di accettabilità p e cioè con probabilità di errore inferiore ad $(1-p)$, tenendo presente la relazione:

$$-t_{1-p} < \frac{m - M}{\frac{s}{\sqrt{n-1}}} < t_{1-p}$$

sarà determinato dalla doppia disuguaglianza:

$$m - t_{1-p}\frac{s}{\sqrt{n-1}} < M < m + t_{1-p}\frac{s}{\sqrt{n-1}}$$

Esempio VII.11

Per applicazione si consideri l'esempio seguente. Si supponga che in una certa provincia sia eseguita un'indagine su un campione di 26 campi coltivati a una certa coltura. La resa media per ettaro ottenuta risulta di 300 Kg con una deviazione standard di 20 Kg. Trovare l'intervallo di confidenza al 95% per il valore medio della resa in tutta la provincia.

In questo esempio M non è conosciuto; inoltre, si hanno i seguenti dati iniziali: $n = 26$, $m = 300$, $s = 20$ ed infine $v = 25$. Per cui sarà $\sigma_m = s/\sqrt{(n-1)} = 20/\sqrt{25} = 20/5 = 4$. Si sa inoltre che, avendo un piccolo campione, la t segue una distribuzione avente un numero di gradi di libertà pari a $v = n-1 = 25$.

Ora, con una probabilità del 95% si avrà:

$$-t_{0,05} < \frac{m - M}{\frac{s}{\sqrt{n-1}}} < t_{0,05}$$

dove $t_{0,05}$ indica il valore di t, con $(n-1)$ gradi di libertà, tale che la probabilità di avere un valore $t > t_{0,05}$ è di appena il 5%.

Allora, l'intervallo di confidenza è determinato dalla seguente relazione:

$$m - t_{0,05}\frac{s}{\sqrt{n-1}} < M < m + t_{0,05}\frac{s}{\sqrt{n-1}}$$

Sostituendo i valori relativi all'esempio considerato nelle disuguaglianze suddette si ottengono i limiti ricercati. Infatti, dalla tavola della t di Student per $v = 25$ gradi di libertà si trova $t_{0,05} = 1,708$ e quindi i limiti cercati saranno:

$$300 - 1,708x4 < M < 300 + 1,708x4$$

$$293,2 < M < 306,8.$$

3.3 *Significatività della differenza tra due medie nei p. c.*

Se si considera la distribuzione della differenza tra le medie di due piccoli campioni, si dimostra che la quantità:

$$t = \frac{(m_1 - m_2) - (M_1 - M_2)}{\sqrt{n_1.s_1^2 + n_2.s_2^2}}\sqrt{\frac{n_1.n_2.(n_1 + n_2 - 2)}{n_1 + n_2}}$$

ha una distribuzione come la t di Student con $v = (n_1 + n_2 - 2)$ gradi di libertà.

Qui l'attenzione deve essere richiamata nei riguardi del fatto che *per usare questa espressione è necessario supporre che le varianze delle due popolazioni siano uguali tra loro*. Ora, nell'applicare questa formula per verificare la significatività della differenza tra le medie dei due campioni si verifica l'ipotesi che sia $M_1 = M_2$. Ciò è fatto ponendo il valore $(M_1 - M_2) = 0$ nel calcolo del valore di t della formula precedente e quindi cercando, come al solito, la probabilità corrispondente nella tavola della t, con $v = (n_1 + n_2 - 2)$ gradi di libertà.

Evidentemente sfruttando la stessa formula data per la t è possibile anche trovare i limiti di confidenza per la differenza tra le medie delle popolazioni da cui i due campioni sono estratti.

A questo scopo, è sufficiente ricavare dalla tavola della t relativa all'intervallo di confidenza al livello del 5% il valore di t che corrisponde ai gradi di libertà del campione considerato e quindi avremo:

$$-t(\nu) < \frac{(m_1 - m_2) - (M_1 - M_2)}{\sqrt{\frac{n_1 . s_1^2 + n_2 . s_2^2}{\frac{n_1 . n_2 . (n_1 + n_2 - 2)}{n_1 + n_2}}}} < t(\nu)$$

per cui dovrà essere soddisfatta la doppia disuguaglianza:

$$(m_1 - m_2) - t(\nu) . \sqrt{\frac{n_1 . s_1^2 + n_2 . s_2^2}{\frac{n_1 . n_2 . (n_1 + n_2 - 2)}{n_1 + n_2}}} < (M_1 - M_2)$$

$$(M_1 - M_2) < (m_1 - m_2) + t(\nu) . \sqrt{\frac{n_1 . s_1^2 + n_2 . s_2^2}{\frac{n_1 . n_2 . (n_1 + n_2 - 2)}{n_1 + n_2}}}$$

che definisce, appunto, l'intervallo nel quale dovrebbe risultare compresa la differenza tra le medie delle due popolazioni. Un esempio mostrerà con chiarezza l'applicazione di queste formule.

Esempio VI.12

Nel corso di un esperimento furono selezionati a caso due insiemi di piante di grano, il primo contenente 17 piante ed il secondo 13. Le altezze medie della spiga per i due campioni sono risultate uguali a 7,5 e 6,0 centimetri e le loro varianze uguali a 0,75 e 0,9 cm rispettivamente. Trovare al livello del 5% di confidenza se la differenza tra i due campioni di altezze delle spighe di grano è significativa. Si ponga $(M_1 - M_2) = 0$ e si ricordi che in questo esempio si hanno i seguenti dati sperimentali: $n_1 = 17$, $n_2 = 13$, $m_1 = 7,5$, $m_2 = 6,0$, $s^2_1 = 0,75$, $s^2_2 = 0,9$. Applicando la formula per il calcolo della t nei piccoli campioni si avrà:

$$t = \frac{7,5 - 6,0}{\sqrt{17x0,75 + 13x0,9}} . \sqrt{\frac{17x13x(17 + 13 - 2)}{17 + 13}} = \frac{1,5}{\sqrt{12,75 + 11,70}} . \sqrt{206,267} = \frac{1,5x14,36}{4,94469} = 4,356.$$

Cercando nella tavola della t il valore trovato $t = 4,356$ con $\nu = 28$ gradi di libertà, si ha che la probabilità relativa è troppo piccola, poiché una probabilità uguale a 0,01 corrisponde solo ad un valore di $t = 2,76$. Di conseguenza questa differenza tra le medie è fortemente significativa e l'ipotesi che sia $M_1 = M_2$ non è giustificata.

E' anche possibile trovare i limiti di confidenza per la differenza tra le medie delle popolazioni da cui i due campioni sono estratti. Usando la stessa formula della t, si ha:

$$t = \frac{1,5 - (M_1 - M_2)}{4,9} . 14,3 = \frac{1,5 - (M_1 - M_2)}{0,3426}$$

Dalla tavola della t l'intervallo di confidenza al livello del 5% con 28 gradi di libertà corrisponde ad un valore di $t = 2,048$ e quindi si avrà:

$$-2,048 < \frac{1,5 - (M_1 - M_2)}{0,3426} < 2,048$$

ossia

$$0,804 < (M_1 - M_2) < 2,196 .$$

3.4 *L'errore standard di s nei p. c.*

Nel caso dei piccoli campioni il valore di σ nella formula:

$$\sigma_s = \frac{\sigma}{\sqrt{2n}}$$

può essere sostituito dal valore della sua stima definito dalla relazione:

$$\hat{\sigma} = s\sqrt{\frac{n}{n-1}}$$

quindi si ha:

$$\sigma_s = \frac{\hat{\sigma}}{\sqrt{2n}} = \frac{s\sqrt{n}}{\sqrt{2n}\sqrt{n-1}} = \frac{s}{\sqrt{2(n-1)}}$$

Il valore di t definito come segue:

$$t = \frac{s - \sigma}{\frac{s}{\sqrt{2(n-1)}}}$$

ha una distribuzione t di Student con $v = (n - 1)$ gradi di libertà.

Per esempio, si consideri un campione di 9 elementi per il quale si è trovato che la deviazione standard è uguale a $s = 2,4$. Si avrà:

$$\sigma_s = \frac{s}{\sqrt{2(n-1)}} = \frac{2,4}{\sqrt{2(9-1)}} = \frac{2,4}{4} = 0,6.$$

Nella tavola della t di Student, a $v = 8$ gradi di libertà e con limiti di confidenza al 5% corrisponde un valore di $t = 2,306$. Quindi è possibile dire che la deviazione standard della popolazione σ cadrà con il 95% di probabilità tra i limiti seguenti:

$$2,4 - 2,306x0,6 < \sigma < 2,4 + 2,306x0,6$$

$$1,014 < \sigma < 3,786.$$